DOCUMENT RESUME

ED 481 119 TM 035 283

AUTHOR	Gorard, Stephen; Smith, Emma
TITLE	Written Evidence for the Inquiry into Secondary Education: Student Achievement.
PUB DATE	2003-00-00
NOTE	16p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Achievement Tests; Foreign Countries; *Secondary Education; *Standards; *Test Use
IDENTIFIERS	*United Kingdom

ABSTRACT

The validity of public examinations as measures of academic achievement is not perfect, and the generalizability of paper-and-pencil tests to real-life tasks is rather low. In the United Kingdom, small differences between levels of attainment on public examinations cannot be attributed to real differences in achievement. The comparability of achievement tests is reduced by changes over time, place, examination board, mode of examining, subject, and syllabus. A major thrust of this paper is to suggest that a consideration of standards or effectiveness is not a simple matter of counting and comparing. In fact, there is no real evidence of failing educational standards over time in Britain and no convincing evidence of underperformance relative to the educational systems of other developed nations. International comparisons and those based on local education agencies do suggest that comprehensive systems of schools based on parental choice tend to produce narrower social differences in intake and outcomes. Systems with more differentiation have greater gaps in attainment between social groups. The United Kingdom is in a reasonable comparative position. There are problems related to education, certainly, but the current examination system was designed to differentiate between candidates. This differentiation cannot be used, logically, as evidence of underattainment. (Contains 54 references.) (SLD)

ED 481 119

Written evidence for the Inquiry into Secondary Education: Student Achievement

Professor Stephen Gorard and Dr Emma Smith
Cardiff University School of Social Sciences
Glamorgan Building
King Edward VII Avenue
CF10 3WT, UK
email: gorard@cardiff.ac.uk

Summary

- 'Achievement' at school generally describes levels of attainment in public examinations such as GCSE.
- The validity of public examinations as measures of achievement is not perfect. The generalisability of pencil and paper tests to real-life tasks is rather low.
- Public examinations are not wholly reliable. Therefore, small differences between levels of attainment cannot be attributed to real differences in achievement.
- Fair and rigorous comparisons cannot be made between different forms of attainment. Comparability is reduced by changes over time, place, exam board, mode of examining, subject and syllabus.
- Differences in attainment cannot be calculated by simple subtraction. They must be proportionate, contextualised, and hedged around with doubts about the underlying distribution of the scores.
- 'Underachievement' is used to describe a range of phenomena. These range from the differential attainment of groups of school students (such as those formed by nation, region, ethnicity, language, school type, sex and social class) to the failure of an individual student to attain a level equivalent to the best prediction of their future performance (value-added or contextualised).
- There are problems of unreliability and invalidity in the categories frequently used to define groups of underachievers (such as social class and ethnicity). As the unreliability of attainment measures and classifying variables increases so does the chance of spurious 'effects'.
- Once operationalised, there is no convincing evidence for any of these forms of underachievement.
- In the UK, there is an absence of appropriate experiments to assess the reasons why some groups do less well in compulsory schooling. Only experimental designs can test causal models leading to fruitful ameliorative action. Filling this gap was the main purpose of the thirty million pounds spent on the ESRC-controlled Teaching and Learning Research Programme. This purpose is unlikely to be met.
- Given this *lacuna*, we are left with *post hoc* analyses of large datasets seeking cause by statistical manipulation, and small-scale studies of 'qualitative' data often not seeking causes at all. Both approaches have significant defects. This paper focuses on the former approach, but the problems generally encountered in the latter approach are even greater in terms of rigour, generalisability and comprehensibility.

- There is no reason to assume that achievement in the UK is worse than in comparable nations. Nor is there any evidence for the much-cited notion that results in the UK are more polarised.
- There is no reason to assume that achievement in different parts of the UK, or in different types of schools, are different for equivalent students.
- There is no reason to assume that achievement differs between social groups, as defined by ethnicity, social class, language or sex (for otherwise equivalent students).
- The differences in raw-score attainment in the above groups disappear in either a value-added or a contextualised analysis.
- There is some evidence that achievement in state-funded schools is improving over time, and that, contrary to popular reports, the gaps in attainment between identifiable groups are declining.
- Much public money is being spent on research that cannot produce the answers required of it, and on policies to ameliorate growing gaps in attainment that do not exist.

There is insufficient space here to argue each of the above closely with full supporting evidence. Instead the outline below uses references to published peer-reviewed material available upon request to supplement the examples of research given.

1. Examinations and comparability

1.1 There have long been complaints that standards of attainment in UK education have fallen over time (Cresswell and Gubb 1990, National Commission on Education 1993, Barber 1996), that they are poor in comparison to similar countries (Boyson 1975, Prais 1990, Skills and Enterprise Network 1999), and that standards are particularly poor for the lowest achievers (Postlethwaite 1985, Bentley 1998, DfES 2001). Therefore, the UK is supposed to have a uniquely polarised assessment system, with excellent results for some and a long tail of underachievers. Claims such as these are quite common, and contribute to what has become a 'crisis account' of the state of the UK education system and its schools (Gorard 2000a).

1.2 However, judging standards is difficult without having a close definition of the term 'standard'. As an illustration of how elastic the term can be, consider the very real situation in which an educational attainment indicator such as a GCSE becomes more common over a period of ten years. One group of comentators may claim that standards have therefore improved, because more students now attain the GCSE standard. Their opponents may claim that standards have fallen, since the GCSE is now demonstrably easier to obtain and also worth less in exchange. The point to be made here is that knowledge is not a static commodity, and comparisons of changes over time in school attainment have to try and take these changes into account. One analogy for the complaint by the National Commission on Education (1993) that number skills have deteriorated for 11-15 year olds, would be the clear drop over the last millennium in archery standards among the general population. If the number of children knowing the meaning of this word 'mannequin' drops from 1950s to the 1970s is this evidence of some kind of decline in schooling? Perhaps it is simply evidence that words and number skills have changed in their everyday relevance. On the other hand, if the items in any test are changed to reflect these changes in society, then how do we know that the test is of the same level of difficulty as its predecessor? In public examinations, by and large, we have until now relied on

norm-referencing. That is, two tests are declared equivalent in difficulty if the same proportion of matched candidates obtain each graded result on both tests. The assumption is made that actual standards of each annual cohort are equivalent, and it is these that are used to benchmark the assessment. How then can we measure changes in standards over time (for there cannot be any, by definition)? But, if the test is not norm-referenced how can we tell that apparent changes over time are not simply evidence of differentially demanding tests? This apparently insuperable problem has, to my mind, not been adequately addressed (Gorard 2001a).

1.3 Britain uses different regional authorities (local examination boards) to examine what are meant to be national assessments at 16+ and 18+ (Noah and Eckstein 1992). It is already clear that even qualifications with the same name (e.g. GCSE History) are not equivalent in terms of subject content as each board sets its own syllabus. Nor are they equivalent in the form of assessment, or the weighting between components such as coursework and multiple-choice. Nor is there any evidence that the different subjects added together to form aggregate benchmarks are equivalent in difficulty to each other. In fact, comparability can be considered between boards in any subject, the years in a subject/board combination, the subjects in one board, and the alternative syllabuses in any board and subject. All of these are very difficult to determine, especially as exams are neither accurate nor particularly reliable in what they measure (Nuttall 1979). The system of statutory assessment is also producing a flood of complaints about irregularities and inconsistencies (Cassidy 1999). Pencil-and-paper tests have little generalisable validity, and their link to other measures such as occupational competence is generally very small (Nuttall 1987).

1.4 The problems faced by researchers in international studies of student performance are even greater. These include the comparability of different assessments, the comparability of the same assessments over time, using examinations or tests as indicators of performance at all, the different curricula in different countries, the different standards of record-keeping in different countries, and the competitiveness (especially) of developing countries (see O'Malley 1998). Yet what international comparisons seek to do is solve not one but *all*, and more of these problems at once (Gorard 2000b).

1.5 A further problem is that simple differences between attainment scores are being routinely misrepresented by academics, policy-makers and the media, in a way that takes no account of their underlying distribution or their base rate (Gorard 1999a, Gorard and Taylor 2002b).

1.6 In summary, it is extremely difficult to claim that small differences in 'surface' attainment between students represent real differences in achievement.

2. Underachievement

2.1 'Underachievement' is now a widely used term in education policy and practice (Gorard 2000c). It is used routinely to refer to nations, home nations and regions, to types and sectors of schooling, to physiological, ethnic and social groups, and to individuals. It has been used to mean simply low achievement, also lower achievement relative to another of these groups, and lower achievement than would be expected by an observer. These multiple uses lead to considerable confusion which, coupled with common errors in assessing the proportionate

difference between groups, mean that significant public money has been spent attempting to overcome problems that may not, indeed, exist (Gorard et al. 2001). Where underachievement is understood to mean a lower level of achievement by an individual (or group) than would be expected using a model based on the *best* available predictors, then the underachieving individuals have nothing in common (else that common factor would become part of the best prediction). If, instead, we reserve some predictors from our best model (sex or poverty, for example), we still find no evidence that underachievers have much in common (Smith 2002). In raw-score terms, we might say that a particular social group exhibits lower achievement (in the sense of publicly available figures relating to pencil and paper tests) than another, as in the case of some ethnic groups. Or we might say that there is differential attainment between groups, as in the case of males and females. This is very far from saying that the lower-attaining group could and should do better on *that* assessment. The term underachievement has conceptual and practical difficulties, which chiefly lie in determining what the 'under' is in relation to. When it is used in relation to peers, or prior attainment, or cognitive aptitude tests for example there is no clear way of separating it from errors in the baseline testing system. To assume, as the DfES and many researchers in this field appear to, that the assessment system is neutral (by sex, for example) and that any differential is related to achievement or performance seems peculiarly naive. This is especially so in the light of the already acknowledged general unreliability of statutory assessments. Making explicit what we mean by underachievement is an important step towards accepting that, collectively, we may not really mean anything by it.

2.2 The nature of formal assessments means that comparing standards over time (or between groups) is very difficult. If the same test is administered repeatedly year-on-year, so that we can assume the same level of difficulty over time, then there are potential practice effects. Any increase in test results could be due to familiarity with the test. On the other hand, where the test is changed every year to keep it up-to-date and prevent practice effects, then we have no way of knowing whether successive tests are of the same standard. Until 1987 this problem was largely overcome in public examinations by 'norm-referencing'. An assumption was made that the test cohort every year was of the same ability, but that the test varied. So, instead of having a pass mark the test had a set pass proportion. For example, in O-level English perhaps 10% were given the top grade every year. So, by definition, it was impossible to ask whether standards were rising year-on-year. The underlying assumption of exam marking was that standards did not change. The only change allowable was in the proportion of the age cohort entering any examination. Since 1987 the UK has moved to a system based largely on criterion referencing. Now, each grade is related to a description of what is required, and if the candidate gives evidence of this then the grade is awarded. Since 1987, therefore, standards have been allowed to vary. This has led to an annual increase in exam scores, but has also made it impossible to tell whether this is due to rising standards of candidates or a lowering of the standards of tests. In the absence of a valid independent benchmark, any discussion of relative educational standards in the UK is somewhat pointless.

National achievement

2.3 Similar problems arise when trying to compare results between countries. Here, the problems of different entry rates and different standardisation procedures are compounded by the different assessment systems, and even by differences in the educational systems (and, of course, the curricula) themselves. Where the same test is administered in each country (as in the Third International Mathematics and Science Study), re-consideration of the results shows that there is

no convincing evidence of 'underachievement' in the UK. UK scores are compared with countries like: the US which has much fuller coverage of the curriculum underlying the test; Singapore where children do not advance through school years automatically (meaning that they were, on average, 6 months older than UK students in TIMSS); and even Thailand whose scores are based only on the 32% of the age cohort attending school. Where a different test is used for each country (perhaps more appropriate to the local curriculum), problems of comparability arise. How can we tell whether the baccalaureate in France or the arbitur in Germany are equivalent in difficulty to the GCSE in the UK?

2.4 Anyway, sixteenth place for England in TIMMS (Mathematics) is far from impressive, but better than several countries including USA, Norway and Spain. Many of the other countries taking part also scored lower, but were omitted by the researchers from analysis as they did not meet the sampling requirements for the study. In this study of the attainment of 14 year-olds, one South American country submitted scores for a cohort averaging 16 years of age. Otherwise, the oldest average age is for Singapore at the top of the table in terms of score, and the youngest is for Iceland near the bottom. The linear correlation between age and score means that one would expect countries with older children in the test to have higher scores, and that nearly 30% of the variance in outcomes is explicable by differences in mean age alone. There are further problems with the study in terms of sampling, low response rate (below 50% for England, Keys et al. 1996), inclusion or exclusion of students with special educational needs, overlap of standard errors, and motivation. Brown (1998) concludes that the information in international league tables is generally too flawed to be of any use at all.

School achievement

2.5 At the level of comparison between schools (department or teachers), school effectiveness work has attempted to describe the characteristics of a successful school in a way that could form the basis of a blueprint for school improvement. Ironically, the major undisputed outcome of all of this work has been the reinforcement of the importance of non-school context (Coleman et al. 1966, Gray and Wilcox 1995). National systems, school sectors, schools, departments and teachers combined have been found to explain approximately zero to 20% of the total variance in school outcomes. In all studies this 'effect' is small, and the larger the sample used, the weaker is the evidence of any effect at all (Shipman 1997) – and, of course, we could not be certain that it is an 'effect' since the underlying causal model remains opaque. The remainder of the variance in outcomes is explained by student background, prior attainment and error components. Despite this, most educational policies are based upon comparisons between schools that do not take these incontrovertible findings into account. Such policies include league tables of results, programmes of inspection, and national and regional targets, all of which have presented attainments in raw-score forms. When researchers have attempted to relate this small school-effect to school characteristics and processes, so producing a blueprint for school improvement, the results have generally been negligible. The factors making up a 'good' school are frequently nebulous (Ouston 1998) or tautological (Hamilton 1997).

2.6 Where claims have been made regarding the superiority of schools in one or more home countries of the UK, the situation is somewhat easier to assess as the systems themselves are more similar. While both countries have very similar school systems, Wales, for example, has until recently produced lower exam scores at all levels than England. However, once levels of poverty have been taken into account, schools in Wales have produced results that are as least as

good as those in England (Gorard 1998a). Similar points can be made about differences between types of schools within one home country (Gorard 1998b). To expect a school with many students in poverty to gain the same kind of exam success as a school with nearly no poor students at all, is ridiculous. Yet this is what raw-score comparisons (such as league tables) do. Once levels of poverty, and other background factors, are taken into account in regression equations then there is no evidence that any type of school performs any better than any other. State-funded schools in the UK are also rapidly catching up with the exam scores of fee-paying schools (Gorard and Taylor 2002b). So the question is not about the underachievement of schools or regions. Rather it is why there is this link between poverty and attainment.

2.7 Once their context is taken into account, there appear to be better and worse performing schools of all types and in all sectors. However, the overwhelming majority of variance in school results is predicted by the nature (or prior attainment) of the intake. Little variance is left to be labelled a 'school effect', and even this contains an error component of unknown size. Put another way, there is no clear evidence of schools having much systematic effect *at all* on the attainment of their students. It appears that each individual would achieve pretty much as they do in any school, and that school 'improvement' consists largely of admitting more high achieving students - whether through direct selection as in some specialist and all grammar schools, or indirectly via the admissions systems, as in faith-based and Foundation schools.

2.8 Operationalising the concept of underachievement is key to appreciating which group of students succeeds at school and also in understanding the confusion between low achievement and underachievement. A recent study used detailed student-level data to measure and identify underachievement among a group of over 2000 year 9 secondary school students (Smith 2002). Over 30 variables which the academic literature cite as being linked to academic performance (such as prior attainment, attitude towards school and receipt of free school meals) were used to predict the future examination performance of these students; any individuals who failed to fulfil their potential were considered to be underachieving. There was little to distinguish the underachieving students from their peers. While there were some working class boys who underachieved, for example, using this definition, there were others who overachieved. Indeed, students in the underachieving group came from across the ability range; therefore it was possible to have a high ability underachiever as well as a low ability underachiever. The best predictors of academic success were prior attainment and attendance at school (accounting for three quarters of the variation in examination outcome), with sex and social class accounting for a negligible amount of the variance. Students who came from more economically disadvantaged backgrounds, performed less well in the Key Stages 2 and 3 examinations in every subject, as well as being less regular attenders at school – disadvantages which far exceeded those between the sexes. However, these students were not disproportionately *under*achieving in terms of the model.

2.9 In summary, once the issues discussed in section 1 are taken on board it is difficult to conclude that levels of attainment in the UK are poor, falling, or weak in comparison to other countries. It is difficult to conclude that any one sector or type of school is weaker than another. It is not possible to identify entire groups of students with a tendency to underachieve. It is possible to identify groups which attain lower scores – but the category which binds them together (such as sex or social class) is a 'pseudo-explanation' for their lower achievement (see below). There is some evidence that standards of attainment are improving over time.

3. Achievement gaps

3.1 This section examines patterns of attainment polarisation in England at a variety of levels. The PISA study in 2000 involved all EU countries. National segregation by examination outcome (for reading – the only score with complete coverage) is largely explicable by the use of academic (and other forms of) selection (Smith and Gorard 2002a). In all countries there are small gaps between the performance of boys and girls in reading, in favour of girls. This gap is generally smaller in countries with the highest overall scores. Overall, the Scandinavian countries of Sweden, Finland and Denmark show less segregation on all indicators. The UK has below average segregation in terms of all indicators, despite a commonly held but unfounded view that segregation in the UK is among the worst in the world.

3.2 Table 1 presents the results for reading performance according to the students' score on the PISA indicator of wealth (Smith and Gorard 2002b). Students who fall into the lowest 10% by wealth perform less well on the reading tests. In general, countries with the lowest gap in reading performance between richest and poorest are also those that have relatively high scores, even for the poorest 10%. Finland, Ireland and the Netherlands have high scores for both groups, while France, Germany and Luxembourg with heavily selective systems have both very low scores for the poorest 10% and only average scores for the richest 90%. The UK has the fourth highest score for the poorest 10% and the third highest score for the richest 90%. In fact, the scores in the UK are so far from polarised that the reading score for the lowest 10% is *higher* than the overall score for most countries. There is no evidence here of the purported crisis of underachievement in UK education. However, all of the foregoing caveats also apply to these figures.

Table 1 - Mean reading score according to PISA indicator of family wealth

Country	Poorest 10%	Richest 90%
Luxembourg	385	452
Portugal	422	483
Germany	454	504
Greece	456	475
France	465	509
Spain	469	499
Italy	472	492
Austria	477	502
Denmark	479	502
Belgium	489	519
Sweden	495	519
UK	502	529
Ireland	512	530
Finland	540	550
Netherlands	541	543

Gaps between groups

3.3 Policy-makers, media commentators, and academics have recently worked together to create a 'moral panic' about the underachievement of boys at school (see for example, DENI 1997, Dean 1998). Although each account may have minor variations, the dominant version is as follows. There was a fairly recent period when boys were out-performing, or at least out-scoring, girls at school. Then girls began to catch up in terms of school performance and qualifications. They have now overtaken the boys, and the gap between the genders is increasing over time. Boys are prevalent in terms of school failure, non-qualification, exclusion and special needs. This is a universal phenomenon unrelated to local socio-economic considerations. Boys are therefore underachieving (see Salisbury et al. 1999 for a fuller account of this literature). Since this much is apparently clear the next task is to overcome the disadvantage of boys by remedial action in schools. This task is being attempted by multiple action research projects (e.g. School of Education 1998) or by attempting to transfer strategies from schools presumed to show good practice because they have a lower gender gap in attainment than their peers (as in the DfES project on 'boys underachievement' based in Cambridge).

3.4 In fact, very little of this dominant account has any validity. The confusion in this field can be seen in the fact that as late as 1997, some respected writers in this field still believed that boys were outscoring girls at GCSE (e.g. David et al. 1997), but that there was 'a closing gender performance gap in most subjects in GCSE' (p.99) with 'girls closing the gender gap' (p.102). Recent re-analyses of the national figures for attainment from Key Stage 1 to A level have shown that the gaps between girls and boys have remained the same since the early 1990s, perhaps even declining slightly over time (Gorard et al. 1999). Where achievement gaps exist (and of the core subjects these only consistently appear in English, and Welsh in Wales), they are at the highest levels of attainment, just as they are for gaps between the achievement of ethnic groups (Johnston and Viadero 2000). The nature and size of these gaps vary regionally, and are clearly related to socio-economic factors. In fact, once the complexity of factors and obstacles such as home background, school structure, and social skills are taken into account a simple gendered explanation of achievement does not work (Kutnick 2000). Nor, apparently, do the simplistic solutions being suggested to the problem, such as single-sex teaching (see Harker 2000). According to the best records we have boys have not attained higher grades (at 16+) than girls for at least 25 years. In fact, it is not even clear that we have any reliable evidence that boys have ever done better than girls in compulsory schooling.

3.5 There is currently no sizeable or consistent gender gap at the lowest level of attainment in any public examination for any subject for any Key Stage. Approximately the same proportions of boys and girls of the relevant age gain at least the lowest level of each qualification (such as Level 1 at Key Stage One). In addition, for Mathematics and Science (and a few other curriculum areas) there is no sizeable or consistent gender gap at any level of attainment. Put another way, the assessment system is *largely* gender-neutral. There are achievement gaps in several curriculum areas, most notably English, other languages, and humanities. Where these appear, they are greatest at the highest level of attainment, mostly affecting a minority of (the most able) children (Table 1). These gaps are not increasing over time. The gaps in some subjects remain relatively static, while some are declining slightly. It is also worth noting that in subjects where children are assessed both by teachers and by a task/test, then the task/test produces lower achievement gaps (i.e. it is more gender neutral).

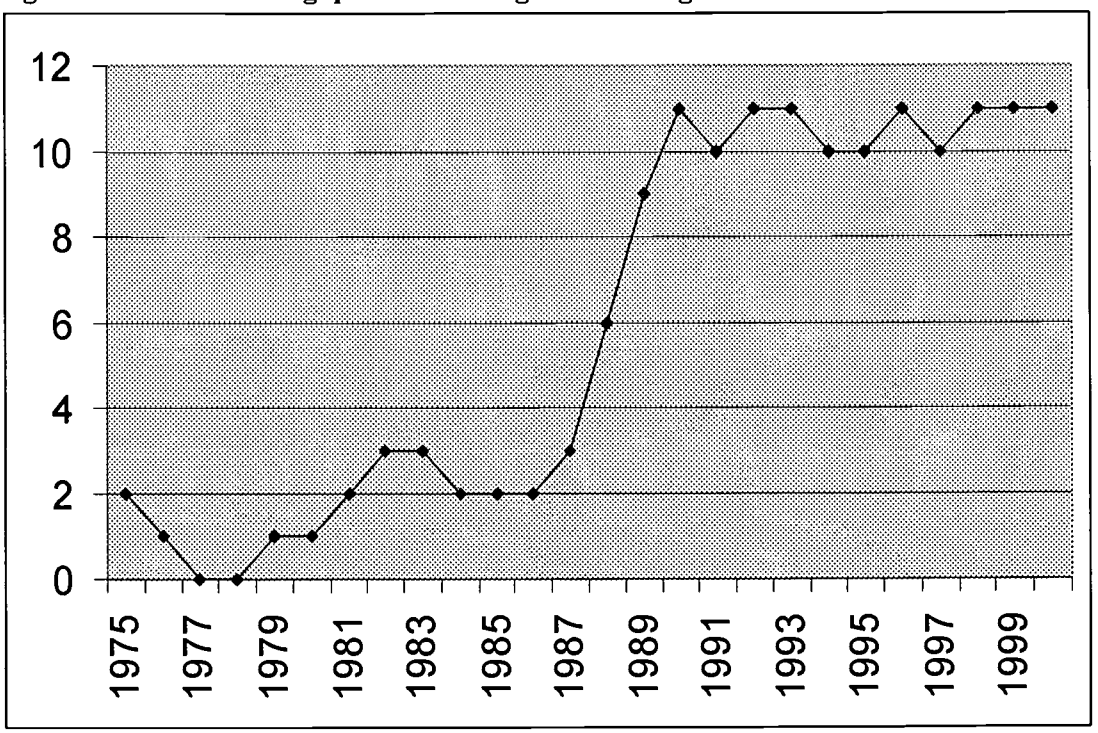
Table 1 - Achievement gap in favour of girls: GCSE English

	Entry	A*	A	B	C	D	E	F	G
1992	20		27	23	16	10	5	1	0
1993	20		31	24	16	10	5	2	0
1994	30	43	34	27	18	11	5	1	0
1995	10	44	35	24	16	8	4	1	0
1996	10	43	36	25	16	9	4	1	0
1997	20	43	35	25	15	9	5	2	1

[table entries represent the extent to which girls outnumber boys in each cell. 20 as an entry gap shows that 20% more girls sit the assessment. 16 as a gap at C grade shows that 16% more girls attain a C or above]

3.6 Figure 2 shows that there are year-on-year fluctuations in the overall achievement gap (in favour of girls), but that girls have never scored lower than boys since 1974 (using the same kind of figures as in the table above). It also suggests that until 1987/88 the overall trend of the gap was relatively static with a low in 1978 and high in 1983. Just at the period when overall scores begin to rise, there is a sudden jump in the size of the achievement gap over a two year period until the gap stabilises again from 1988/89 to 1997/98. This could explain the perplexing, to some, finding that from 1992 to 1997 the gender gap in a subject-by-subject analysis remained constant (Gorard 2001b). In summary, the gender gap at GCSE is chiefly a phenomenon appearing between 1987 and 1989, and growing only during that same period. This information could be key to our understanding of the determinants of this gap.

Figure 2 - Achievement gap in favour of girls attaining 5+ GCSE A*-C



3.7 The differential attainment of boys and girls at 16+ has appeared over a relatively brief period since 1987, concurrent with major increases in qualification levels for the entire 16-year-old cohort. The introduction of the GCSE heralded several other major changes including the abolition of strict norm-referencing at O level which had previously worked to maintain results at a relatively constant level (Foxman 1997). This was linked to the largest ever annual increase in the proportion of those reaching the GCSE (or O level equivalent) benchmark in 1988, and the second largest in 1989. In addition, the publication of the results for the 16-year-old cohort replaced the previous School Leavers Survey (which had included results from children of other age-groups) and formed the basis for new school performance tables. It is surely no coincidence that the gender gap appeared at precisely the same time as these changes, along with the introduction of course-work assessment and the onset of the National Curriculum with SATs, which according to evidence from the Youth Cohort Study have all greatly increased the chances of success for those from 'poor' backgrounds (Dolton et al. 1999).

3.8 The potential practical importance of such a basic finding cannot be over-estimated. Given that the gender gap is, in fact, related to both social class and levels attainment, then the appearance of a large gap just when children from poorer families began to score more highly begins to suggest possible explanations. Consider the following as one example of an implication for ameliorative strategies. If the notion of what constitutes 'work' and what is appropriate for home varies by occupational class, it may be that 'working-class' men, and their boys, do not bring work home, whereas 'working-class' women, and their girls, do. If so, strategies such as homework clubs or Saturday sessions at school may be more natural and therefore effective for such boys than homework pacts. Another possible conclusion to be drawn from this would be that differential attainment by sex is a product of the changed system and nature of assessments rather than any more general failing of boys, their ability, application, or the competence of those who teach them. Such a conclusion, that differences are highly dependent on the nature of assessment, would be supported by the recent debate over the apparent improvement in boys' literacy as a result of the literacy hour where sensitivity to the precise nature of the test appeared to determine the nature of the gender gap (Cassidy 2000), and by the finding that achievement gaps can vary considerably depending on whether the assessment is by teacher or task/test.

3.9 Similar findings apply, where data are available, to differential attainment by ethnic group and by economic region. It is not clear why differences between ethnic groups, regions, and genders occupy so much commentator attention. The gaps between other social groups, such as by first language or between rich and poor, are much larger than the gender gap. Perhaps the biggest single gap is between the high and low achievers. The achievement gaps between the top and bottom 10% are very large, and completely dwarf any differences between boys and girls. However, these gaps are also inherent in the nature of the assessment system. A system that did not differentiate at all would be dismissed, but it is clearly possible to change the assessment system to reduce the gap in 'surface' attainment between any groups.

3.10 Although the methods used here allow fair comparisons over time and place, there is no method suitable for comparing gaps in tests scores between different age groups. It is impossible, for example, to decide whether the gender gap is larger, smaller or the same at Key Stage Four as it is at Key Stage Two (although this does not prevent commentators from making spurious comparisons based on expected levels). The metrics are not equivalent. However, the gender gap in qualifications, such as it is, reverses among adults in later life.

4. Implications

4.1 One thrust of this paper has been to suggest that a consideration of standards or effectiveness is not a simple matter of counting and comparison (Gorard 2000d). Even where simplifying assumptions are made about the outcomes from schools, such as a concentration on statutory assessment and test results, philosophical and methodological difficulties persist. In light of these difficulties, there is certainly no evidence here of falling educational standards over time in Britain, no convincing evidence of underperformance relative to the educational systems of other developed nations, and no evidence of a highly polarised system.

4.2 International and LEA-based comparisons do suggest that comprehensive systems of schools based on parental choice tend to produce narrower social differences in intake and outcomes. Systems with more differentiation lead to greater gaps in attainment between social groups. Finland, for example, has a high average reading score, a small gap between high and low attainers and comprehensive schools and a policy of choice. Germany, on the other hand, has a much lower average reading score, a large difference between high and low attainers, and a tiered system of selective schooling. The UK is currently still in a reasonable comparative position, with a high average reading score, below average differences between high and low attainers, and comprehensive schools with a policy of (limited) choice. The lessons for current policy are obvious.

4.3 However, not all commentators are aware of this. There is a common crisis account of the position of UK schooling (and there is, perhaps, a tendency for all commentators to decry the position of their own countries). For example, Johnson (2002) recently complained that 'British students may be among the world's highest achievers, as the recent Organisation for Economic Co-operation and Development's PISA study found. But the achievement gap between social classes remains one of the biggest in the world' (p.23). This represents the view of the IPPR – an influential centre-left think tank. The introduction of choice policies have, according to this account, led to a greater polarisation of results. But this greater polarisation by parental occupation does not exist in the UK. Something that does not exist cannot, therefore, be the result of choice policies.

4.4 There is no particular urgency about the issue of differential attainment by gender (certainly no more so than around 1988 when the only big increase took place) or by any other physiological group, and there may be many hidden dangers in tinkering with a school system already near 'initiative-overload'. Both the action research approach and the transfer of 'successful' school strategies for raising the attainment of boys might therefore be considered both wasteful and unnecessary (not to mention inequitable). Since the current gap has existed since 1988, is not growing, and is much smaller than other systematic gaps (such as those by background of student), we can take our time to search for ameliorative solutions if they are required. It might, for example, be much simpler to obtain gender neutrality through a reconsideration or redesign of the assessment system (whence the gap may have come), than through changes in classroom interaction (and similar comment apply to ethnicity). While still facing potential problems such as teacher supply and inequitable funding arrangements, on any rational comparison the UK school system is in the healthiest state ever. Raw score indicators of attainment are rising annually, gaps between social groups are reducing, and socio-economic

segregation between schools has declined. We do not appear to need yet more major interventions to solve problems that do not exist and that detract from dealing with the problems that do.

4.5 Perhaps the most important conclusions to be drawn are negative ones. The fact that boys and girls perform the same at low levels of attainment (or indeed at all in some subjects), coupled with the relative stasis of the gender gap since 1989, suggests that many potential explanations are now unworkable. Any useful causal explanation would focus on high, not low, level attainment, and suggest an instant one-off impact. Notably therefore, this differential attainment is not the result of a cultural change in society, new methods of teaching, seating arrangements in schools, mixed-sex classes, boys' laddishness, or poor attendance at school. This has serious implications for the conduct of future work, and for the validity of previous work, in this area. Longitudinal work with large-scale datasets has elucidated the overall pattern, while the action research and the transfer-of-successful-strategies approaches adopted by the DfES have been unhelpful at this stage. In fact, a considerable amount of public funding is being wasted in attempting to solve a specific problem of underachievement at school that does not actually exist.

4.6 Another conclusion to be drawn from this would be that differential attainment by gender is a product of the changed system and nature of assessments rather than any more general failing of boys, their ability, application, or the competence of those who teach them. Such a conclusion - that differences are highly dependent on the nature of assessment - would be supported by the recent debate over the apparent improvement in boys' literacy. This improvement was apparently the result of sensitivity to the precise nature of the test. It might, for example, be much simpler to obtain gender neutrality through a reconsideration or redesign of the assessment system (whence the gap may have come), than through changes in classroom interaction. Whatever ameliorative strategies are proposed, it would be preferable for them to be considered carefully in light of a fuller analysis of differential attainment than hitherto (especially through a consideration of the *interaction* of gender, ethnicity, poverty and so on). This should also be done with the full realisation that all such strategies may have longer term impacts on the lives of both men and women in adult society.

4.7 Value-added analyses of individual student performance data have called into question the underachievement of large groups of students.

4.8 The use of school improvement models has led, indirectly, to an overemphasis on the most visible indicators of schooling - examination and test scores. There is a considerable danger of targets, based on these indicators, determining the practice of organisations. The use of test scores leads to three related problems. It may marginalise other purposes and potential benefits of schooling. In addition, it suggests that variations in the scores themselves are the product of school effects when the evidence clearly shows otherwise. It also neglects the fact that the scores themselves are artificial, and technically difficult to compare fairly over time or place. Our current examination system was designed to differentiate between candidates. If it did not do so, it would be rejected, presumably, as ineffective. We cannot, logically, use this differentiation *per se* as evidence of underachievement.

4.9 All of the findings from the kind of studies described here, and smaller-scale studies of classroom processes, will remain open to dispute until a decision is made to demand definitive

experimental testing of the possible determinants of achievement. Meanwhile, we are often left with mere pseudo-explanations such as sex, region or social class. Even if we could show that sex was a *cause* of a level of achievement, we could not adjust the sex of individuals to ameliorate low achievement. Even if we find that achievement varied by region, we would be foolish to believe that transplanting populations between regions would be practical or effective (and so on). This is what we mean by 'pseudo-explanations'.

4.10 General improvements in the standards of, and outcomes from, education appear to be reducing the educational inequalities between different social groups and geographical regions. Kelsall and Kelsall (1974) present some evidence that the gap between the top and bottom of the social scale in economic, power and status terms was being reduced by the 1970s. Although inequality and injustice for the socially disadvantaged has always existed (MacKay 1999), in fact, 'if you take a long-term historical perspective of the provision of education in the UK throughout its entire statutory period... you could say that a constant move towards greater justice and equity has been the hallmark of the whole process' (p.344). If so, good.

References

Barber, M. (1996) *The learning game*, London: Indigo

Bentley, T. (1998) *Learning beyond the classroom*, London: Routledge

Boyson, R. (1975) *The crisis in education*, London: Woburn Press

Brown, M. (1998) The tyranny of the international horse race, pp.33-47 in Slee, R., Weiner, G. and Tomlinson, S. (Eds.) *School Effectiveness for Whom? Challenges to the school effectiveness and school improvement movements*, London: Falmer Press

Cassidy, S. (1999) Test scores did not add up, *Times Educational Supplement*, 16/7/99, p. 5

Cassidy, S. (2000) Keep it short and simple for boys, *Times Educational Supplement*, 14/1/00, p.5

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. and York, R. (1966) *Equality of educational opportunity*, Washington: US Government Printing Office

Cresswell, M. and Gubb, J. (1990) The Second International Maths Study in England and Wales: comparisons between 1981 and 1964, in Moon, B., Isaac, J. and Powney, J. (Eds.) *Judging standards and effectiveness in education*, London: Hodder and Stoughton

David, M., Weiner, G. and Arnot, M. (1997) Strategic feminist research on gender equality and schooling in Britain in the 1990s, pp. 91-105, in Marshall, C. (Ed.) *Feminist critical policy analysis: Volume 1*, London: Falmer Press

Dean, C. (1998) Failing boys public burden number one, *Times Educational Supplement*, 27/11/98, p. 1

DENI (1997) *A review of research evidence on the apparent underachievement of boys*, Department of Education Northern Ireland Research Briefing RB5/97

DfES (2001) *Schools - Achieving Success*, HMSO, London

Dolton, P., Makepeace, G., Hutton, S. and Audas, R. (1999) *Making the grade: Education, the labour market and young people*, York: Joseph Rowntree Organisation

Eurydice (2002) *The Information Network on Education in Europe*, www.eurydice.org [Accessed October 2002]

Foxman, D. (1997) *Educational league tables: for promotion or relegation?*, London: Association of Teachers and Lecturers

Gorard, S. (1998a) Schooled to fail? Revisiting the Welsh school-effect, *Journal of Education Policy*, 13, 1, 115-124
Gorard, S. (1998b) Four errors.... and a conspiracy? The effectiveness of schools in Wales, *Oxford Review of Education*, 24, 4, 459-472
Gorard, S. (1999) Keeping a sense of proportion: the "politician's error" in analysing school outcomes, *British Journal of Educational Studies* , 47, 3, 235-246
Gorard, S., Salisbury, J. and Rees, G. (1999) Reappraising the apparent underachievement of boys at school, *Gender and Education*, 11, 4, 441-454
Gorard, S. (2000a) Questioning the crisis account: a review of evidence for increasing polarisation in schools, *Educational Research*, 42, 3, 309-321
Gorard, S. (2000b) *Education and Social Justice*, Cardiff: University of Wales Press
Gorard, S. (2000c 'Underachievement' is still an ugly word: reconsidering the relative effectiveness of schools in England and Wales, *Journal of Education Policy* , 15, 5, 559-573
Gorard, S. (2000d) One of us cannot be wrong: the paradox of achievement gaps, *British Journal of Sociology of Education*, 21, 3, 391-400
Gorard, S. (2001a) International comparisons of school effectiveness: a second component of the 'crisis account'?, *Comparative Education* , 37, 3, 279-296
Gorard, S. (2001b) An alternative account of 'boys underachievement at school', *Welsh Journal of Education* , 10, 2, 4-14
Gorard, S., Rees, G. and Salisbury, J. (2001) The differential attainment of boys and girls at school: investigating the patterns and their determinants, *British Educational Research Journal* , 27, 2
Gorard, S. and Taylor, C. (2002a) What is segregation? A comparison of measures in terms of strong and weak compositional invariance, *Sociology*, 36, 4, 875-895
Gorard, S. and Taylor, C. (2002b) Market forces and standards in education: a preliminary consideration, *British Journal of Sociology of Education*, 23, 1, 5-18
Gray, J. and Wilcox, B. (1995) *'Good school, bas school' Evaluating performance and encouraging improvement* , Buckingham: Open University Press
Hamilton, D. (1997) Peddling feel-good fictions, in White, J. and Barber, M. (Eds.) *Perspectives on school effectiveness and school improvement*, London: Institute of Education
Harker, R. (2000) Achievement, gender and the single-sex/coed debate, *British Journal of Sociology of Education*, 21, 2, 203-218
Johnson, M. (2002) 'Choice' has failed the poor, *The Times Educational Supplement*, 22nd November, p.23
Johnston, R. and Viadero, D. (2000) Unmet promise: raising minority achievement, *Education Week*, 15/3/00, pp.1-8
Kelsall, R. and Kelsall, H. (1974) *Stratification*, London: Longman
Keys, W., Harris, S. and Fernandes, C. (1996) *Third International Mathematics and Science Study: National Report Appendices*, Slough: NFER
Kutnick, P. (2000) Girls, boys and school achievement, *International Journal of Educational Development*, 20, 1, 65-84
MacKay, T. (1999) Education and the disadvantaged: is there any justice?, *The Psychologist*, 12, 7, 344-349
National Commission on Education (1993) *Learning to succeed*, London: Heinemann
Noah, H. and Eckstein, M. (1992) Comparing secondary school leaving examinations, pp. 3-24 in Eckstein, M. and Noah, H. (Eds.) *Examinations: comparative and international studies*, Oxford: Pergamon Press

Nuttall, D. (1979) The myth of comparability, *Journal of the National Association of Inspectors and Advisers*, 11, 16-18

Nuttall, D. (1987) The validity of assessments, *European Journal of the Psychology of Education*, II, 2, 109-118

O'Malley, B. (1998) Measuring a moving target, *Times Educational Supplement*, 18/9/98, p.22

Ouston, J. (1998) *The school effectiveness and improvement movement: a reflection on its contribution to the development of good schools*, presented at ESRC Redefining Education Management seminar, Open University, 4/6/98

Postlethwaite, T. (1985) The bottom half in secondary schooling, pp. 93-100 in Worswick, G. (Ed.) *Education and economic performance*, London: NIESR

Prais, S. (1990) Mathematical attainments: comparisons of Japanese and English schooling, in Moon, B., Isaac, J. and Powney, J. (Eds.) *Judging standards and effectiveness in education*, London: Hodder and Stoughton

Salisbury, J., Rees, G. and Gorard, S. (1999) Accounting for the differential attainment of boys and girls at school, *School Leadership and Management*, 19, 4, 403-426

School of Education (1998) *Report on combating underachievement, especially in boys: an action research project*, Cardiff: School of Education, Cardiff University

Shipman, M. (1997) *The limitations of social research*, Harlow: Longman

Skills and Enterprise Network (1999) *Skills and Enterprise Network Annual Conference Report*, Sudbury: DfEE Publications

Smith, E. and Gorard, S. (2002a) International equity indicators in education: defending comprehensive schools III, *Forum*, 44, 3, 121-122

Smith, E. and Gorard, S. (2002b) *What does PISA tell us about equity in education systems?*, Occasional Paper 54, Cardiff University School of Social Sciences

Smith, E., (2002) *Could do better? Understanding the nature of underachievement at Key Stage 3*, paper presented at the British Education Research Conference, University of Exeter, 11th-14th September 2002

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").